Mach Translat (2016) 30:41–62
DOI 10.1007/s10590-016-9188-5

How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data

Lucas Nunes Vieira[1]

Received: 28 July 2016 / Accepted: 20 December 2016 / Published online: 21 January 2017

Abstract There has been growing interest of late in the cognitive effort required by post-editing of machine translation. Compared to number of editing operations, cognitive (or mental) effort is frequently considered a more decisive indicator of the overall effort expended by post-editors. Estimating cognitive effort is not straightforward, however. Previous studies often triangulate different measures to obtain a consensus, but little post-editing research to date has attempted to show how measures of cognitive effort relate to each other in a multivariate analysis. This paper addresses this by presenting an exploratory comparison of cognitive measures based on eye tracking, pauses, editing time, and subjective ratings collected in a post-editing task carried out by professional and non-professional participants. All measures correlated with each other, but a principal components analysis showed that the measures cluster together in different ways. In particular, measures that increase with task time alone behaved differently from the others, with higher mutual associations and higher reliability. Regarding differences between professional and non-professional participants, it was observed that subjective ratings were overall more strongly associated with objective measures in the case of professionals. Surprising findings from previous research based on pause ratio are discussed. The paper argues that a pause typology will benefit the study of pause lengths and cognitive effort in post-editing.

Keywords Post-editing · Machine translation · Cognitive effort · Eye tracking · Pauses · Subjective ratings

✉ Lucas Nunes Vieira
l.nunesvieira@bristol.ac.uk

[1] School of Modern Languages, University of Bristol, 17 Woodland Road, Bristol BS8 1TE, UK

1 Introduction

The potential benefits of post-editing machine translation (MT) output, as opposed to translating source texts from scratch, are now largely uncontroversial in the context of non-literary translation (cf. Green et al. 2013; Plitt and Masselot 2010). However, post-editing MT is not beneficial on all occasions. MT quality and the effort required by post-editing can be influenced by a number of factors, such as genre/domain and source-text features that are problematic for MT (cf. Bernth and Gdaniec 2001; Calude 2004). In view of this, there has been growing interest in measuring the effort required by post-editing for the purpose of examining the feasibility of this practice and for empirically identifying characteristics of the source text or MT output that can be used as effort predictors (e.g. Aziz et al. 2014). Similarly, there have been studies aimed at identifying the extent to which different individual profiles affect post-editing effort. Some previous research in this respect has found no effect of prior professional experience on post-editing time (e.g. de Almeida 2013; Guerberof 2014), though a more recent study has found that expert post-editors are more productive than novices (Moorkens and O'Brien 2015). A positive attitude to MT is also often found to be a factor in post-editing performance (e.g. de Almeida 2013; Mitchell 2015). Moorkens and O'Brien (2015) observed that attitudes tend to be more negative in the case of professionals.

Most studies mentioned above use editing time and/or number of editing operations as proxies for effort. According to a now recurrent post-editing effort typology proposed by Krings (2001), editing time and number of editing operations reflect the notions of *temporal* and *technical* effort, respectively. Underpinning these two effort dimensions is the notion of *cognitive* (or mental) effort, which, according to Krings, 'involves the type and extent of [the] cognitive processes that must be activated in order to remedy a given deficiency in a machine translation' (ibid.: 179). In other words, cognitive effort is the type of effort related to the mental processes that take place during a task, which are not necessarily proportional to technical effort. This is because cognitive decisions do not necessarily involve any edits (e.g. when the MT output is left in its original state) or indeed because editing operations do not necessarily involve a high level of mental effort; they might just be mechanically laborious. Krings argues that cognitive effort is the central variable that controls the amount of time post-editors spend on the task and guides the modifications they perform (ibid.: 179). It has previously been shown that cognitive effort does not necessarily correlate with the number of changes implemented in the raw MT output (e.g. Koponen 2012).

A variety of measures can be used as proxies for cognitive effort. Previous studies frequently combine these in the hope that the strengths of one method offset the weaknesses of another, but there has been little research to date on the potential overlaps between these measures and on their degrees of reliability. The present paper investigates these issues by presenting an exploratory multivariate comparison of different measures used in previous research to estimate cognitive effort. The purpose of the study is to address questions and generate hypotheses with respect to two specific aims:

– checking to see how subjective ratings, eye-tracking metrics, pauses and editing time relate to each other in a multivariate analysis considering both professional and non-professional participants; and
– checking to see which of these measures present the highest levels of reliability.

In the remainder of the paper, Sect. 2 provides an overview of the concept of cognitive effort and reviews previous research on the topic. Section 3 presents the paper's methodology. Sections 4 and 5 present and discuss results, respectively, and Sect. 6 concludes the paper by providing a summary of findings and directions for future research.

2 Review of literature

Measurements of cognitive effort can serve a variety of purposes, such as estimating the level of difficulty of a task or, in a post-editing scenario, the quality of the MT output. Cognitive effort is an elusive construct, however—both conceptually and methodologically. In cognitive psychology, *effort* has in many contexts been treated as a synonym for *attention*. Kahneman (1973) provided an early account in this respect. He discusses the notion that paying attention to a task is equivalent to allocating mental resources to it (ibid. 13).[1] The amount of resources available, or our mental capacity, is limited, which means that the higher the amount of resources expended (and the closer we get to the resource limit), the higher the amount of effort (cf. Moray 1967). These early notions have been more recently developed in the context of cognitive load theory, which relies on three central concepts: the notion of a load that is imposed on our mental system (i.e. the demands of a task), the notion of the effort expended to cope with this load and the notion that the balance between load and effort will be linked to individuals' performance (Kirschner 2002, p. 4).

In post-editing, mental load can be roughly represented by the difficulty of the task (e.g. the quality of the raw MT output and the level of complexity of the source text). Mental effort is the overarching variable analysed in this paper, i.e. the mental effort necessary to cope with the task demands. Performance relates to how the task is carried out (e.g. the number of errors participants make) (Paas et al. 2003, p. 64). Recent research on cognitive effort in post-editing has primarily focused on contrasting cognitive effort with the task demands (e.g. Lacruz et al. 2014; O'Brien 2006). Among other things, correlations between these two constructs allow post-editing effort to be predicted based on specific characteristics of the source text or the MT output. Since cognitive effort cannot be measured directly, a number of indirect measures have been used in previous research. The present study investigates the behaviour of seven such measures: eye fixation count, average fixation duration, editing seconds per word, subjective ratings, pause ratio, pause-to-word ratio, and average pause ratio (see Sect. 3.4).

[1] The concept of attention is also associated with the human capability of selecting only certain elements to process out of a number of competing stimuli. This facet of attention is different from effort (see Kahneman 1973: 2).

Eye movements have a long tradition in cognitive research. The count and average duration of eye fixations[2] have been used in a variety of studies as proxies for cognitive effort in post-editing (e.g. O'Brien 2011). The rationale for the use of eye tracking to estimate cognitive effort is based on the eye-mind and immediacy assumptions (Just and Carpenter 1980), which posit that by fixating the eyes on the text while reading individuals necessarily, and immediately, engage in the mental processing of the content read. Based on these assumptions, higher fixation duration and count are a sign of more mental processing and therefore more effort.

Subjective ratings have also been traditionally used as cognitive effort estimates. Here, the rationale is that increasing task load produces a sensation of effort that individuals can report in numerical terms (O'Donnell and Eggemeier 1986, p. 7). In post-editing, subjective ratings have been used as a measure of cognitive effort by, for example, Koponen (2012) and Vieira (2014). Moorkens et al. (2015) contrasted subjective ratings with more objective measures, such as eye movements and number of changes in the raw MT output. Moorkens et al. do not provide an indication of how different measures behave in relation to each other in a multivariate analysis, however, constituting a different approach to the one presented here.

Pauses in text production have been used as measures of cognitive effort in a variety of contexts. Typing is usually regarded as the material representation of a given mental process, so pauses in text production are deemed to represent the act of replacing these processes, i.e. by engaging in a new mental process after the previous one has been materialised (see Schilperoord 1996, pp. 8–9). O'Brien (2006) examined a potential correlation between pause ratio in post-editing (i.e. pause time over total task time) and negative translatability indicators (i.e. source-text features that are problematic for MT). Surprisingly, O'Brien did not observe this correlation and suggested that pauses alone are not a robust metric to estimate cognitive effort in post-editing. Interestingly, Daems et al. (2015) found a significant correlation between pause ratio and MT errors. Daems et al. contrast MT errors with a number of other cognitive effort measures, including eye movements and pause metrics, but this study again does not provide an indication of how these measures relate to each other in a multivariate analysis, which makes it hard to identify what the measures could indeed be reflecting, or what they have in common. Lacruz et al. (2012) and Lacruz and Shreve (2014) propose an improvement to pause ratio and argue that average pause ratio (i.e. average pause length over average time per word) and pause-to-word ratio (i.e. number of pauses per word) are more sensitive to clusters of shorter pauses, which are assumed as indicators of cognitive effort in post-editing.

With regard to temporal measures, Koponen et al. (2012) argue that by normalising editing time by target text length, it is possible to capture the amount of mental processing that takes place in post-editing. Based on a cognitive typology of MT errors proposed by Temnikova (2010), Koponen et al. show that the average number of seconds spent per target word in post-editing is higher in sentences containing a larger number of cognitively expensive errors. They show that editing time normalised in this way can be used as a proxy for cognitive effort.

[2] Fixations are defined as 'eye movements that stabilize the retina over a stationary object of interest' (Duchowski 2007, p. 46). That is, when our eyes 'fixate' on an object (e.g. a word or parts of it).

The studies reviewed above show a number of ways in which the cognitive effort expended in post-editing can be estimated. It should be noted, however, that casting a wide net and using a large number of different measures in the hope that together they will provide a more accurate parameter might be an inefficient approach. This is especially the case when the measures are correlated, the likely case when they are all expected to reflect the same construct. Using highly correlated measures as outcome variables (i.e. the variables assumed to depend on others) in a study requires multivariate strategies that can handle all variables at once. This issue merits attention because not only is using correlated outcome variables a redundant approach, but this also increases the number of tests performed, which inflates the chance of false positives and requires some type of correction (see Snijders and Bosker 1999). Furthermore, previous research hardly presents a discussion of what differs between cognitive effort measures or what these measures could in fact be reflecting. Aziz et al. (2014) use a multivariate method to analyse post-editing data, but they analyse correlations between textual features and post-editing effort (editing time and operations), rather than correlations between measures of cognitive effort themselves. The analysis presented here explores this issue and provides a framework that can shed light on the concept of cognitive post-editing effort, allowing educated decisions to be made with regard to cognitive effort estimation in future research.

3 Methodology

The analysis presented in this paper is based on data collected by Vieira (2016) in the context of a larger project. Further methodological details can also be found in Vieira (2014).

3.1 Source text and machine translation output

The source texts were extracts of two online news articles[3] taken from the Workshop on Statistical Machine Translation (WMT) 2013 corpus.[4] The articles were written in French. They were about prostate cancer and the United States elections. English machine translations were sampled from the WMT corpus and from a pool of additional translations harvested for the study from freely available MT engines.[5] It seemed desirable to expose participants to a range of conditions. Harvesting outputs from

[3] See http://www.lapresse.ca/vivre/sante/201211/30/01-4599309-depistage-du-cancer-de-la-prostate-passer-le-test-ou-non.php and http://www.lapresse.ca/la-tribune/opinions/201207/30/01-4560667-une-strategie-republicaine-pour-contrer-la-reelection-dobama.php. Accessed 23 July 2016.

[4] This corpus results from an annual MT shared task where submitted systems are ranked according to a human evaluation. Source texts, reference translations and system submissions are available at http://www.statmt.org/wmt13/results.html. Accessed 15 June 2016.

[5] The WMT systems used in the study were 'Online-A', 'Online-B', 'Online-G', 'CU-Zeman', 'FDA', 'CMU-syntax', 'rmbt-3', 'rmbt-4', 'KIT', 'DCU-FR-EN-Primary', 'MES-Simplified', 'Edinburgh' and 'Edinburgh-unconstrained'. The outputs purposely harvested for the investigation were from SDL Freetranslation.com (http://www.freetranslation.com), TransPerfect (http://web.transperfect.com/free-translations/) and Microsoft Translator; the latter via Microsoft Word. These outputs were harvested in October 2013.

these additional systems helped to achieve this by widening the range of MT quality used in the study. The Meteor automatic MT evaluation system (Denkowski and Lavie 2011) was used to evaluate the MT sentences so that sentences from a range of MT quality levels could be identified and included in the study's sample.

The Meteor scoring system assesses the level of matching between a machine translation and a human reference translation. Meteor scores range from 0 (complete mismatch between machine translation and reference) to 1 (perfect match). Version 1.4 of Meteor was used in the study with default settings. Human translations in the WMT 2013 corpus were used as reference. Sentences at different Meteor levels, produced by different systems, were combined into a single machine-translated text to be presented for editing. This was done by randomly selecting sentences at each available decile of the Meteor spectrum (for further details see Vieira 2014). Sentences in the study sample had Meteor scores ranging between 0.14 and 1.

In total, 1037 source words were presented for editing. Data corresponding to the title of the articles and to sentences without reference translations were excluded from the analysis, leaving 844 words (41 sentences) available for the study of cognitive effort, based on the two texts combined. The absence of reference translations meant that Meteor scores could not be computed for these sentences. Excluding the titles, in turn, seemed desirable as a way of avoiding 'acclimatisation effects'—i.e. when more effort is spent while participants are still familiarising themselves with the task/text (cf. Doherty et al. 2010, p. 9).

3.2 Post-editing task

Participants post-edited the two texts in two sessions, with a break in between. The task was carried out in PET (post-editing tool) (Aziz et al. 2012), on a computer connected to a Tobii 120Hz eye tracker.[6] The order of presentation of the texts was alternated between participants. Each text was nevertheless presented for editing in document order. As in most studies on post-editing and translation, this means that observations for each sentence were not independent from each other, as participants' behaviour in editing one sentence could have affected their behaviour in editing subsequent ones (e.g. because of repeated words later in the text, which might require less effort to be processed the second time round). However, since the present study is aimed at comparing the effort measures themselves, rather than checking to see how they associate with a specific condition, this is not regarded here as particularly problematic: all measures compared were subject to the same interdependence between the sentences.

One sentence pair (source-target) was shown on screen at a time and participants were not allowed to backtrack. While these operating conditions restrict the study's findings to the sentence level, these conditions were necessary for a more reliable collection of the gaze data. The eye-tracking data was processed by establishing the source-target sentence pairs as an area of interest (AOI) in PET's interface and extracting the fixations landing on this area with Tobii Studio. The source and target sentences

[6] With the Toby I-VT fixation filter (Olsen 2012), set to filter out individual fixations below 100ms. The other filter settings available were kept at default values.

were both included in the AOI as both these elements are expected to involve cognitive effort (e.g. in understanding and interpreting the source text, and in identifying and correcting problems in the MT output).

Participants were not allowed to consult external sources, but before each session they read short factual descriptions of the topics covered in each text, which was aimed at reducing the impact of the external sources restriction. The task had no time limit, but a post-editing brief was given to participants before the task with the recommendation to aim for a high-quality post-edited product in as little time as possible.

After post-editing both texts, participants took working memory capacity and source-language (French) proficiency tests, and filled in a final questionnaire. Working memory capacity is defined as 'the amount [of information] that an individual can hold in mind at one time' (Cowan 2005, p. 3). This is expected to affect the expenditure of cognitive effort in post-editing because individuals with a higher capacity of holding information in the mind might find it easier to correct certain types of errors, for example those involving long textual spans (cf. Temnikova 2010). An automatic reading span task (Unsworth et al. 2005, 2009) was used to measure working memory capacity in the study. The French test, in turn, involved identifying real French words among plausible non-words (Meara and Buxton 1987), a type of task that is deemed reliable for linguistic placement purposes (cf. Read 2007). In the present study, results from the French test were used as a way of filtering out participants with a low level of source-language proficiency. Information on participants' professional and academic background, and their attitude towards MT, found by previous research to be a factor in editing performance (de Almeida 2013), was collected in the final questionnaire. Attitude to MT was rated on a scale between 1 (negative) and 5 (positive).

3.3 Participants

The tasks were conducted on Newcastle University's campus. The involvement of human participants in the study was authorised by the University's ethics committee. The sample included participants who were students, freelance translators, or both (e.g. postgraduate students with professional translating experience). The sample in Vieira (2016) included participants with varying levels of professional experience and proficiency in French, variables that were controlled for in the analysis carried out therein. Statistically controlling for participant variables is not straightforward with some the methods used here (e.g. principal components analysis), however, so it seemed desirable to subset the original sample and make use of a more balanced group of participants. This was done by selecting equal numbers of professionals and non-professionals among those who were above the original sample's average of French proficiency. This selection resulted in a group of ten participants. Details of their profile are presented below in Table 1.

All participants were English native speakers. The analysis presented here is contrasted between participants with professional experience (>0.1 year) and without (≤ 0.1 year), with five participants in each group. Mann-Whitney tests confirmed that the groups are comparable with respect to French knowledge, age, attitude to MT, and

Table 1 Participants' profile

Participant	FR Vocab (0–100)	Experience (in years)	Age	Attitude to MT (1–5)	Working memory capacity (0–75)
P01	79	1.5	23	5	25
P05	95	0	55	5	36
P07	97	0.1	22	4	44
P08	95	0	21	4	68
P09	97	0	22	4	61
P10	89	4	26	3	38
P13	95	3	60	2	37
P15	93	0	20	3	46
P16	93	5	37	3	33
P18	97	4	30	2	50
Mean	93	1.7	31.6	3.5	43.8

working memory capacity,[7] while a significant difference exists in terms of years of professional experience ($W = 0$, $p = 0.01$). Ultimately, variability in the characteristics presented in Table 1 is expected to exist in real post-editing settings, so this helps to guarantee that the sample covers a realistic range of post-editor profiles.

3.4 Cognitive effort measures

The post-editing process measures analysed here are those used in previous research as indicators of cognitive effort. All measures were computed at a sentence level. A brief description of these measures is provided below.

Fixations per word (FPW): this measure consists of the total number of fixations normalised by source word count to avoid sentence-length effects.

Average fixation duration (AFD): this measure was obtained by dividing total fixation time by the total number of fixations. In addition to being analysed in its own right, this measure was used as a parameter for screening the quality of eye-tracking data. As per previous research (cf. O'Brien 2011), data points where average fixation duration was lower than 200 ms were excluded (7% of the data).

Pause-to-word ratio (PWR): this is a cognitive effort measure proposed by Lacruz and Shreve (2014) (see Sect. 2). Pause data was obtained from PET key-log files by calculating the intervals between any keyboard or mouse activity that resulted in a change in the text (navigation events were disregarded). This is similar to the strategy followed by Carl and Kay (2011) and Green et al. (2013). As per Lacruz et al. (2014), only when these intervals were 300 ms long or more they were considered pauses. The pause-to-word ratio measure was obtained by dividing the total number of pauses by the number of source words.

[7] Age: $W = 5.5$, $p = 0.1995$; French: $W = 14.5$, $p = 0.661$; working memory capacity: $W = 19$, $p = 0.1658$; attitude to MT: $W = 17.5$, $p = 0.2714$.

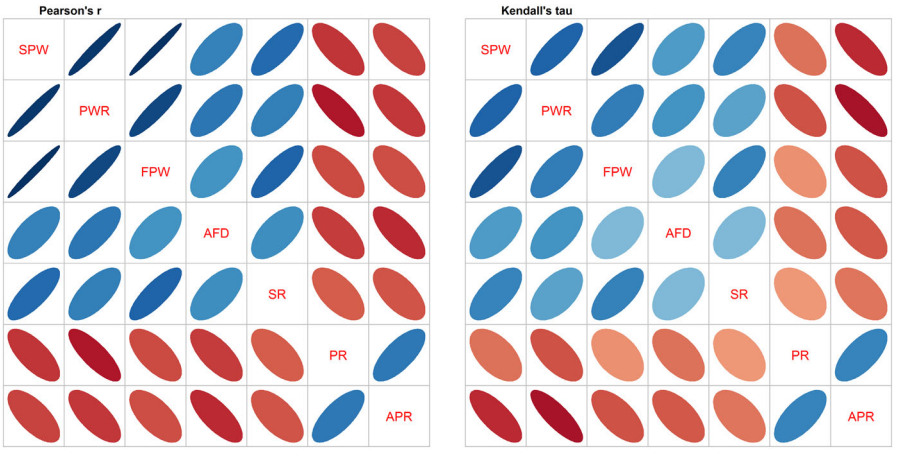

Fig. 1 Correlation matrix of cognitive effort measures based on Pearson's r (*left*) and Kendall's tau (*right*). Narrower ellipses indicate stronger correlations, and the ellipses' orientation indicates the correlation direction. All cells display significant correlations ($p < 0.05$), with the Holm correction for multiple tests

Average pause ratio (APR): this measure was obtained by dividing the average time per pause (i.e. average pause length) by the average time per source word (see Lacruz et al. 2012).
Pause ratio (PR): this measure was obtained by dividing the total pause time in a sentence by the total time spent post-editing that sentence (cf. O'Brien 2006). The total post-editing time per sentence was obtained from PET log files.
Seconds per word (SPW): as per Koponen et al. (2012), this measure was computed by dividing total post-editing time by the target (i.e. post-edited) word count (see Sect. 2).
Subjective ratings (SR): this measure was based on a scale largely used in educational psychology to measure 'the perceived intensity of mental effort' (Paas 1992, p. 429). The scale ranges between 1 ('very, very low mental effort') and 9 ('very, very high mental effort') (ibid: 430). It was set up in PET's interface with internal levels (2–8) unlabelled. Participants were prompted to choose a level of the scale on screen after confirming each sentence. These ratings are treated as a numerical variable in the study.

4 Results

4.1 Comparing different measures

The average total editing time for the task was 34.3 minutes (range 21–45.2). To provide an overview of how the measures described in Sect 3.4 are associated with each other, Fig.1 presents the correlation matrix of the measures based on per-sentence averages.[8]

[8] The average of all participants was taken per sentence to prevent observations corresponding to the same participant or sentence being treated as independent measurements. This procedure was followed in Sects. 4.1 and 4.3.

Narrower ellipses in Fig.1 show stronger correlations and the ellipses' orientation shows the effect's direction (leaning forward = positive; leaning back = negative). All measures of cognitive effort analysed were found to correlate with each other. This is the expected behaviour of these measures, as they are all used as proxies for cognitive effort. However, Fig.1 shows considerable variation in the effects' strength, indicating that these measures might cluster together in different ways. In particular, seconds per word (SPW), pause-to-word ratio (PWR) and fixations per word (FPW) have stronger correlations between themselves when compared to the other measures. Correlations involving average fixation duration (AFD), pause ratio (PR), and subjective ratings (SR) are weaker overall. It is also noteworthy that PR and APR have negative correlations with all other measures, which is consistent with the rationale proposed by Lacruz et al. (2012), where higher APR values indicate lower cognitive effort. This negative effect is explored in detail in Sect 4.3 with respect to PR and the association between cognitive effort and pause lengths in post-editing.

Figure 1 shows that not all measures of cognitive effort have the same type of relationship with each other. However, various pairwise correlations are expected to involve a great degree of redundancy, which hinders a precise examination of what these measures have in common and what they do not. Principal components analysis (PCA) (Jolliffe 2002) was used here as a way of addressing this. Informally, PCA transforms a group of variables into a group of orthogonal principal components (PC) containing linear combinations of the original variables. The PCs incrementally maximise the original variables' variance, which means that a small number of PCs is usually enough to explain most of the original data. It should be noted that this method is not sensitive to non-linearity. While non-linear dimensionality reduction techniques are available—e.g. t-distributed stochastic neighbour embedding (van der Maaten and Hinton 2008)—this paper wishes to obtain a visualisation of how the measures relate to each other, rather than a visualisation of how the data points are distributed in the space. PCA is a deterministic and straightforward method that can be used to visualise networks of measures in this way. In addition, when used for descriptive purposes, PCA has 'no need for explicit distributional assumptions' (Jolliffe 2002, p. 19).

A correlation matrix PCA of the cognitive effort measures described in Sect 3.4 was carried out with the `princomp` R function for the purpose of PC selection. Baayen (2008, p. 121) suggests a rule of thumb that regards as important only PCs accounting for at least 5% of the variance. This was the case of the first three PCs, which had 77, 9.3 and 5.4% of the variance, respectively. These results indicate that the measures do roughly agree with each other, as they loaded mostly onto a single PC. However, in line with Baayen's 5% rule of thumb, the second and third PCs are also used here for further observations. These PCs could hold information about slightly different aspects of cognitive effort that are not accounted for by the 77% of variance explained by the first PC. This is consistent with the exploratory nature of the study, where a maximum variance approach seems desirable.

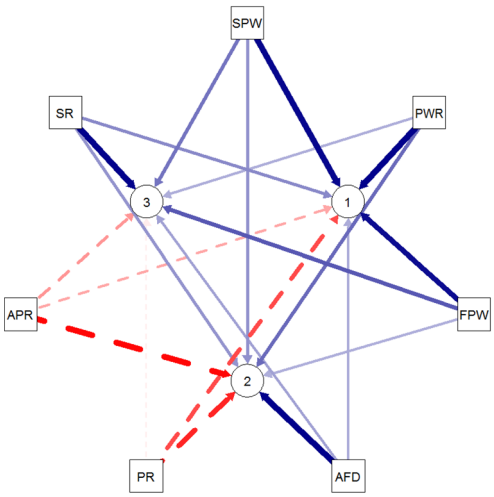

Fig. 2 Plot showing relationship between PCs and cognitive effort measures

To visualise how the measures map onto the first three PCs, Fig. 2 shows a plot generated with the `qgraph.pca` R function[9] (Epskamp et al. 2012) illustrating how the cognitive effort measures and the PCs interconnect. The three circles at the centre of Fig. 2 represent the first three PCs and the square nodes around the circles are the cognitive effort measures. The arrows indicate what measures are loaded onto each PC and with what strength; thicker lines indicate stronger loadings and dashed lines indicate negative loadings. Figure 2 shows that seconds per word (SPW), pause-to-word ratio (PWR) and fixations per word (FPW) loaded more strongly onto PC1, while average pause ratio (APR), pause ratio (PR) and average fixation duration (AFD) loaded more strongly onto PC2. APR and PR have a negative relationship with the other measures, as indicated by the dashed lines (and as shown in Fig. 1). Subjective ratings (SR) is the only measure with a strong loading onto PC3, indicating that this PC discriminates between SR and all the other measures.

To check how these relationships might differ between professional and non-professional participants, correlation matrix PCAs were carried out separately for the two groups. Again using the `princomp` R function, only PCs with a minimum of 5% of the variance were considered. This included four PCs in both cases, with 72, 11, 6, and 5% of the variance for professionals, respectively; and 72, 10, 7, and 7% of the variance for non-professionals. Plots obtained with the `qgraph.pca` R function are presented in Fig. 3.

Figure 3 shows that not all measures relate to each other in the same way when professional and non-professional participants are compared. A particularly interesting difference between the two groups is the behaviour of SR. When subjective ratings on cognitive effort were provided by professional participants, this measure loaded

[9] The option `rotation="varimax"` was used for plots generated with this function throughout the paper. The varimax rotation is a procedure that 'maximizes the sum of the variances of the squared loadings' in the selected PCs (see Dunteman 1989, p. 49), a procedure that makes it easier to observe what differs between them.

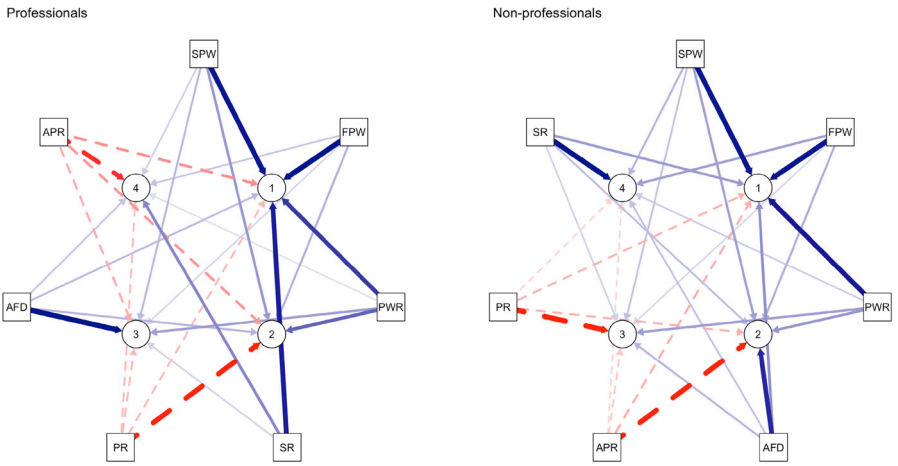

Fig. 3 Plots showing relationship between PCs and cognitive effort measures for professionals (*left*) and non-professionals (*right*)

more strongly onto PC1, together with SPW, FPW and PWR. These results suggest that professional participants might be more inclined to fusing temporal aspects of the task (e.g. SPW) with subjective cognitive effort. Differences can also be noted in the behaviour of APR and AFD between the two groups; these measures had a stronger mutual association in the case of non-professionals. SPW, PWR and FPW, on the other hand, had a very similar relationship for both professionals and non-professionals, with stronger loadings onto PC1 in both cases, which applies particularly to SPW and FPW.

4.2 Measurement reliability

The results presented so far provide an idea of the ways in which the measures analysed are associated with each other. Another aspect to these results is the extent to which these variables are reliable across individual post-editors. Information in this respect is important as this may dictate the best measures to be chosen in experiments where only a small number of post-editors is available. To investigate this, the intraclass correlation coefficient (ICC) of the measures was calculated. This score works as a measure of inter-rater reliability, giving an indication of the extent to which different post-editors produce the same values of a given measure when exposed to the same items. While results presented in Sect. 4.1 are based on averages of all participants, the ICC scores are based on participants' individual scores for each sentence. Here, whenever a sentence needed to be excluded for a given participant due to low-quality eye-tracking data (see Sect. 3.4), the sentence was excluded for all participants so that only complete cases without any missing data remained (24 sentences).

Table 2 shows two-way random[10] ICCs in descending order, together with 95% confidence intervals obtained after 10,000 bootstrap resamples. ICC scores normally range

[10] That is, where subjects and sentences are assumed to be sampled from a population.

Table 2 Intraclass correlation coefficient (ICC) for cognitive effort measures with bias-corrected and accelerated (BCa) bootstrapped confidence intervals (95%)

	ICC	BCa lower bound	BCa upper bound
Pause-to-word ratio (PWR)	0.60	0.53	0.68
Seconds per word (SPW)	0.55	0.47	0.64
Fixations per word (FPW)	0.50	0.41	0.60
Pause ratio (PR)	0.42	0.30	0.56
Average pause ratio (APR)	0.35	0.18	0.44
Subjective ratings (SR)	0.27	0.17	0.40
Average fixation duration (AFD)	0.25	0.17	0.36

between 0 (chance agreement) and 1 (perfect agreement). As can be seen in Table 2, the measure with the highest agreement across participants was PWR, followed by SPW, FPW, PR, APR, SR, and AFD. These results suggest that in contexts with a small number of post-editors, PWR would provide the most reliable estimates of all measures analysed, though the difference in relation to SPW, the measure with the second best result, was not found to be significant according to the 95% confidence interval.

Interestingly, the ranking in Table 2 shows that measures that loaded more strongly onto PC1 in the PCAs presented above (PWR, SPW, and FPW—see Figs. 2 and 3) are more reliable than the others. This further suggests that the overall behaviour of these measures can be distinguished: PWR, SPW, and FPW have higher mutual correlations and are more reliable, while the other measures have lower correlations and lower reliability. It is noteworthy that these differences are not due just to a matter of method. FPW and AFD, for example, two measures based on eye tracking, had strikingly different results both in terms of how they relate to other measures and in terms of their reliability scores. It is also noteworthy that AFD had the lowest ICC result despite being a traditional measure of cognitive effort with a research tradition that goes beyond post-editing and translation studies. In this respect, it is worth pointing out that the fact that a given measure is less reliable does not mean that this measure is less capable of reflecting the underlying construct being estimated. Rather, this might simply mean that a larger number of participants is required to provide a more reliable consensus in the case of studies interested in more objective parameters.

4.3 Pause ratio and cognitive effort

The fact that pause ratio (PR) presented an inverse relationship with cognitive effort (i.e. a negative correlation with other measures) is arguably surprising. This has also been pointed out by Daems et al. (2015), who observed a similar effect in examining the correlation between PR and MT errors. As mentioned in Sect. 2, the rationale for the use of pauses as indicators of cognitive effort in text production is that pauses are deemed to represent the time required for replacing processes in the mind before materialising (e.g. typing) the text. Not necessarily mental processing will only occur during pauses, but based on work by Butterworth (1980) the general assumption made

in previous research is that 'longer pauses reflect cognitive processes that are relatively more effortful compared to processes reflected by shorter pauses' [emphasis removed] (Schilperoord 1996, p. 11).[11] This suggests that if pauses are longer and take up a larger proportion of editing time (i.e. higher PR), this would be expected to indicate higher and not lower cognitive effort. The opposite was observed here, however; both the average length of pauses and their ratio to editing time (i.e. PR) – measures that were correlated based on Kendall's test ($r_\tau = 0.68, p < 0.001$)—were found to have a negative relationship with other measures of cognitive effort.

Lacruz et al. (2012) show that high pause densities indicate high cognitive effort in post-editing. This places more emphasis on the number of pauses than on their length and, to an extent, is consistent with the PR effect observed here. If the data is ranked according to PR, the top 25% quantile of sentences (where PR is higher) has 1.7 pause lasting 11.3 s, on average, whereas the bottom 25% quantile of sentences (where PR is lower) has 20.8 pauses lasting 2.7 s, on average, indicating that higher pause count (and lower PR) is associated with higher cognitive effort. This difference between the quantiles suggests that it is more logical to regard PR and APR as indirect indicators of pause count, rather than measures that are capable of reflecting the relationship between the length of pauses and cognitive effort. This is consistent with the suggestion made by Lacruz and Shreve (2014) that pause-to-word ratio (a measure of pause count) works as a substitute for APR.

The fact that cognitive effort has a clearer relationship with pause count than with pause length has at least two implications: pause length may not constitute a useful parameter for effort estimation in post-editing, or it may be that the relationship between pause length and cognitive effort varies according to different pause types. It is argued below that the latter possibility seems more plausible.

It was observed that a number of data points in the present study (21%) had PR values of 1, which was also observed by O'Brien (2006). These data points represent occasions where no modification is performed in the MT output—i.e. when all the time spent on a sentence consists of a single pause, which divided by itself gives a value of 1. Leaving the MT output unedited is expected to require a certain degree of cognitive effort, as some level of cognitive processing will be involved in making the decision that the text does not require intervention.[12] It seems plausible however that the level of cognitive processing and effort associated with decisions of this kind will be lower compared to decisions made in the context of problem-solving events, where the MT output needs to be edited. Based on a review of the literature on the use of pauses in translation research, Kumpulainen (2015) suggests, for example, that pauses can indicate both problematic and unproblematic processing, and that 'problem spots seem to require more cognitive effort' (2015, p. 55).

Pauses that correspond to the process of leaving the MT output unedited, in particular, seem like good examples of pauses reflecting unproblematic processing. However, the PR measure does not discriminate between different pause types; it assumes that

[11] Schilperoord's study is based on dictated texts, but he argues that there is no evidence to suggest that this significantly affects this assumption (1996, pp. 20–23).

[12] It may also be that some of these decisions are at least partially automatic, but an in-depth investigation of automaticity and decision-making in post-editing is beyond the scope of the paper.

Table 3 Correlations of reduced versions of pause ratio (PR1 and PR2) and corresponding average pause length (APL1 and APL2) with FPW, SR and AFD

	Pearson's r				Kendall's tau			
	PR1	APL1	PR2	APL2	PR1	APL1	PR2	APL2
Fixations per word (FPW)	0.70**	0.42*	0.51**	−0.08	0.60**	0.33	0.33	−0.03
Subjective ratings (SR)	0.81**	0.61**	0.68**	0.15	0.69**	0.40	0.41	0.11
Average fixation duration (AFD)	0.68**	0.30	0.53**	−0.20	0.49*	0.13	0.25	−0.16

P values (** 0.01; * 0.05) have been adjusted for multiple tests based on the Holm method

the length of a pause consisting just of reading the MT output and leaving it unedited has the same weight as the length of a pause that occurs in between typing events, in a problem-solving process. It is hypothesised here that this lack of discrimination between problematic and unproblematic processes could be one of the reasons for the *negative* association observed here between PR and cognitive effort.

As a first step in exploring this possibility, the PR measure was recalculated whilst tentatively filtering out pauses that were not directly associated with problem-solving events. Excluding all such pauses would require a detailed classification of the cognitive processes that take place during each pause, which is not straightforward based on current methods (cf. O'Brien 2006; Kumpulainen 2015). An approximation can nevertheless be achieved by excluding pauses that are not both preceded and followed by typing events in the process of working through a sentence. In the present study, these are (a) pauses that consist of just reading the sentence and leaving it unedited, and (b) the first and last pauses that take place when editing a sentence. Pauses of type (b) may involve both unproblematic and problematic processes. For example, the first pause that occurs when editing a sentence may involve the reading that takes place before a problem is spotted and the process of spotting a given problem and thinking of an initial solution. The last pause may involve checking if the solution for a specific problem has worked or simply reading the sentence again in search for other potential problems, without any being found. Since it is not straightforward to tease these processes apart, for exploratory purposes two modified versions of PR were used here: one where pause types (a) and (b) both assume a value of 0—henceforth 'pause ratio 1' (PR1); and one where only pause type (a) assumes a value of 0—henceforth 'pause ratio 2' (PR2).

Table 3 shows how PR1 and PR2 and the corresponding average pause length (APL1 and APL2) correlate with FPW, AFD and SR, measures that loaded onto different PCs as per the analysis shown in Fig. 2. As can be seen, not all correlations were found to be significant and they have different strengths. However, it is noteworthy that PR acquires a positive association with other measures of cognitive effort after filtering out pauses expected to involve unproblematic processing, in line with the expectation that longer pause length corresponds to higher cognitive effort. This positive association was only found to be significant based on both Pearson's and Kendall's tests in the case of PR1, i.e. when pause types (a) and (b) both assumed a value of 0. This suggests that these pauses can substantially influence the behaviour of PR in post-editing.

It should be noted that Lacruz et al.'s (2012) assumption that pause clusters indicate higher cognitive effort was confirmed by the present data. Based on this assumption, pause types (a) and (b) should not be excluded as this is unlikely to significantly affect the relative pause counts (i.e. because either one or two pauses would be excluded for all sentences). All pauses that take place in a task are in principle of interest depending on the study's objectives and on what is defined as a pause (see Kumpulainen 2015). Nevertheless, results in Table 3 serve to illustrate that simply summing up the length of all pauses to calculate pause ratio in post-editing may lead to surprising negative effects that contradict previous assumptions on the link between pause lengths and cognitive effort.

Previous translation research has noted a similar problem. Immonen and Mäkisalo (2010, p. 49), for example, only consider pauses occurring in fluent target production sequences 'to avoid the skewing effect that revision may have on pause length'.[13] Daems et al. (2015) imply something similar. In view of the longer durations of pauses occurring at the beginning of segments in their study, they propose that it might be worth examining 'pauses in isolation' (2015, p. 43). Indeed, based on the results presented above it is argued here that post-editing studies interested in pause length might benefit from a typology that discriminates between different pauses and the ways in which the lengths of each pause type might relate to cognitive effort. This might allow aspects relating to more passive moments of the task to be studied in greater detail; for example the distribution of post-editors' attention during pause types (a) and (b), and the driving factors of under-editing, when problems in the MT output are left unedited.

To demonstrate how the process of filtering out pause types (a) and (b) affects the relationship between PR and other cognitive effort measures in a multivariate analysis, PCAs were carried out on the measures described in Sect. 3.4 together with PR1, the only measure found to have a significant association with the others as per both correlation tests presented in Table 3. The PCAs were carried out separately for professional and non-professional participants.

Figure 4 shows the results obtained with the `qgraph.pca` R function based on the first three PCs, which before rotation, based on the `princomp` R function, had 70, 10 and 7% of the original measures' variance for professionals, and 77, 8 and 5% for non-professionals.[14] As can be seen, PR1 has a positive association with all other measures except PR and APR. In addition, it is noteworthy that PR1 and SR have strong loadings onto PC3, suggesting a strong connection between these measures. This effect is stronger for non-professionals, but for both participant groups subjective ratings were more strongly associated with the modified version of PR (i.e. PR1) than with PR itself. This suggests that when participants provide subjective ratings on mental effort, they tend not to associate pause types (a) and (b) with effortful moments of the task.

[13] Immonen and Mäkisalo (2010) compared pauses in translation and monolingual writing, and concluded that in both activities pause durations relfect the size of the linguistic unit processed.

[14] The first four PCs would be chosen for professionals according to Baayen's (2008) 5% rule of thumb, but the first three were selected in this case for comparability with the group of non-professionals.

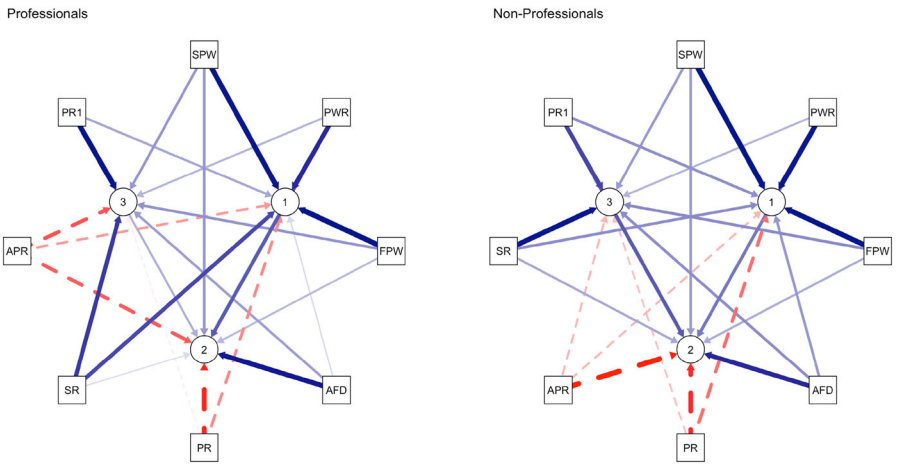

Fig. 4 Plots for professional (*left*) and non-professional participants (*right*) showing relationship between cognitive effort measures including PR1

5 Discussion

The results presented above showed that PWR (pause-to-word ratio), SR (subjective ratings), AFD (average fixation duration), FPW (fixations per word), SPW (seconds per word), and PR (pause ratio), measures used in previous research to estimate cognitive effort, all correlate with each other. However, these correlations varied in strength and direction, indicating that certain combinations of measures have a higher amount of redundancy. A PCA carried out on averages of all participants showed that SPW, PWR and FPW loaded more strongly onto the same PC. These measures also presented a higher level of reliability compared to the others. SR was among the measures presenting the most salient differences between professional and non-professional participants: subjective ratings provided by professionals had a stronger association with more objective measures such as FPW. Moorkens et al. (2015) contrast subjective ratings with temporal effort for professionals and students, but their results are not directly comparable to the ones obtained here as the rating categories used in their study ask participants to rate the amount of editing the MT output is expected to require, rather than cognitive or mental effort itself.

Overall, results obtained in the present study suggest that measures of cognitive effort used in post-editing have different behaviours and can be distinguished with respect to their level of reliability and the ways in which they cluster together. At least two reasons could be behind these differences. First, it is plausible that all these measures are susceptible to a certain degree of measurement error, with variations in this respect potentially driving a distinction between them based on different levels of noise/precision. A second possibility is that different measures may be more sensitive to different nuances of cognitive effort, which would imply that, while a single construct, cognitive effort might have different facets. This second possibility seems like a more plausible explanation for the clustering patterns observed here than just

methodological imprecision, since measures obtained via the same method (e.g. FPW and AFD) had different results in terms of their reliability and mutual association.[15]

In practical terms, a relatively straightforward distinction that can be made between the measures analysed here is the extent to which they are expected to increase with task time alone. This is a particular characteristic of SPW and FPW, for example, where given an MT sentence of a certain length, the longer post-editors take to edit the sentence, the more seconds and fixations the task will involve, i.e. higher values of SPW and FPW. To a certain extent, this is also expected of PWR, since longer tasks are expected to involve more pauses. This would not be expected of AFD, APR, PR, and SR, however. In the case of AFD, for example, not necessarily post-editors will have a longer average fixation on a sentence just by taking longer to post-edit it; they might simply have many short fixations on that sentence, which will increase FPW but not AFD. As SPW, FPW and PWR presented a different behaviour in relation to the other measures, it may be that a capability of indicating temporal aspects of cognitive processing is one of the phenomena underpinning these results.

Measures that are not directly influenced by total editing time, by contrast, are perhaps more sensitive to a pure notion of cognitive effort that can be empirically distinguished from temporal effort. Given the lower reliability scores for measures such as AFD and APR, it also seems that measures that are not directly influenced by total task time are more sensitive to participant variation. This would be consistent with the notion that cognitive effort is an inherently subjective variable that depends on how *individuals* cope with variations in the demands of a task (see Sect. 2), an aspect that future research could explore in more detail.

Regarding the direction of the associations observed here, it was noted that PR and APR presented negative correlations with the other measures. This is consistent with the rationale proposed by Lacruz et al. (2012), who show that the APR measure, in particular, is more sensitive to pause density, which indicates higher cognitive effort in post-editing. In terms of pause length, however, the inverse association between cogntive effort and the PR measure seems to contradict the assumption that longer pauses reflect higher levels of effort in text production (Schilperoord 1996; Butterworth 1980). It was shown here that by disregarding certain post-editing pauses expected to reflect mostly unproblematic cognitive processes, PR acquired a positive association with cognitive effort. Filtering out these pauses was an approximate operation carried out for exploratory purposes, and is not suggested here as a standard procedure. In fact, especially in the case of PWR, Lacruz et al.'s (2012) assumption that higher pause density reflects higher cognitive effort seems quite robust irrespective of pause types, as PWR had the highest level of reliability of the measures analysed in the present study. However, what the results on PR obtained here suggest is that the study of pause lengths in post-editing might benefit from a typology that allows different pause types to be identified and analysed in their own right. This is related to an observation made by Schilperoord (1996, p. 20) regarding an inherent weakness of using key-log files to study pauses in writing. He highlights, for example, that 'the writer may have been reading the text on screen, or he [sic] may have been thinking about what to write next,

[15] On the differences between FPW and AFD and the potential multifaceted nature of cognitive effort, see also Doherty et al. (2010) and van Gog et al. (2009).

but [in key-logging] there is no way of distinguishing such differences'. It was found here that this affects the behaviour of PR, so triangulating methods (see Kumpulainen 2015) with a view to documenting different pause types in post-editing seems like an interesting direction for future research on the relationship between pause lengths and post-editing cognitive effort.

6 Concluding remarks

6.1 Summary and conclusion

This paper aimed to compare cognitive effort measures in a multivariate analysis, and to explore potential differences between the measures in terms of measurement reliability and in how they related to each other considering professional and non-professional participants. Although the study's sample is relatively small and limited to a single language pair, the results obtained suggest that all measures of cognitive effort considered are associated with each other. However, these associations have different strengths. Exploratory analyses showed that the measures cluster together in different ways. Fixations per word, seconds per word and pause-to-word ratio, in particular, had a higher mutual association and higher levels of reliability relative to average pause ratio, average fixation duration, pause ratio and subjective ratings. The behaviour of subjective ratings was found to be different between professional and non-professional participants. While these participant subsets were small, this measure had a stronger association with more objective parameters, such as seconds and fixations per word, in the case of professionals.

These results are capable of informing a number of methodological decisions in future post-editing research. The correlation tests and principal components analyses carried out here indicate that certain measures of cognitive effort are strongly correlated with each other whereas others are less so. This should arguably be taken into account in future studies as a way of avoiding redundancy and carrying out fewer statistical tests, which in large number inflate the chance of false positives. With regard to reliability, the results obtained here suggest that studies with few participants and which are interested in more objective parameters should avoid using measures at the bottom of the ranking displayed in Table 3, such as subjective ratings and average fixation duration. For studies interested in individual variations in cognitive effort, on the other hand, these measures seem like more sensitive parameters.

It was postulated that one of the reasons underlying the clustering patterns observed here is the extent to which the measures are influenced by total editing time. Measures observed to have higher mutual associations and higher reliability (e.g. seconds and fixations per word) are expected to increase with task time alone and might represent the notion of temporal effort more directly, reflecting cognitive effort by extension. Measures that do not increase with task time alone (e.g. average fixation duration), by contrast, are proposed here to reflect a purer notion of cognitive effort that might be empirically distinguished from task time.

It was also observed that the length of pauses that do not occur in between editing events, such as the first and last pauses for a sentence and pauses related to leaving the MT output unedited, influences the behaviour of pause ratio. When this measure included such pauses, it had an inverse association with cognitive effort. When these pauses assumed a value of zero, the association between pause ratio and cognitive effort became positive. Furthermore, it was found that participants do not seem to associate pauses that do not occur in between editing events with effortful moments of task, since the version of PR that excluded these pauses was more strongly associated with participants' subjective ratings. In view of these results, this paper suggests that post-editing comprises different types of pauses and that not all of these will have the same kind of relationship with cognitive effort in terms of their length.

6.2 Directions for future research

Future research could examine the behaviour of other measures of cognitive effort that could not be investigated here, ideally by contrasting sentence-level and discourse-level analyses. Saccades in eye-tracking data and the amount of crossing that occurs between the source and target text, in particular, might be interesting measures to take into account. Pupil dilation and EEG (electroencephalogram) data are other indicators of cognitive processing that are worth examining. The analysis presented here does not consider non-linearity, so this is another aspect that future research could examine in exploring these relationships. In view of the results presented above, looking into the concept of temporal effort and the extent to which it influences the behaviour of different measures seems like an interesting question to be further analysed. There also seems to be room for further work on the impacts of individual profiles on post-editing behaviour. Previous professional experience was selected here as a factor of interest. Future research could look into how factors such as conscientiousness and different types of training might affect cognitive effort in post-editing.

References

Aziz W, Castilho S, Specia L (2012) PET: a tool for post-editing and assessing machine translation. In: Proceedings of LREC 2012, eighth international conference on language resources and evaluation, 21–27 May 2012. Istanbul, Turkey, pp 3982–3987

Aziz W, Koponen M, Specia L (2014) Sub-sentence level analysis of machine translation post-editing effort. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 170–199

Baayen RH (2008) Analysing linguistic data: a practical introduction to statistics using R. Cambridge University Press, Cambridge

Bernth A, Gdaniec C (2001) MTranslatability. Mach Transl 16:175–218

Butterworth B (1980) Evidence from pauses in speech. In: Butterworth B (ed) Language production. Speech and talk, vol 1. Academic Press, London, pp 155–176

Calude AS (2004) Machine translation of various text genres. Te Reo N Z Linguist Soc J 46:67–94

Carl M, Kay M (2011) Gazing and typing activities during translation: a comparative study of translation units of professional and student translators. Meta 56:952–975
Cowan N (2005) Working memory capacity. Psychology Press, New York
Daems J, Vandepitte S, Hartsuiker R, Macken L (2015) The impact of machine translation error types on post-editing effort indicators. In: Proceedings of the fourth workshop on post-editing technology and practice (WPTP4), MT Summit XV Oct 30–Nov 3 2015. Miami, FL, USA, pp 31–45
de Almeida G (2013) Translating the post-editor: an investigation of post-editing changes and correlations with professional experience. PhD Thesis, Dublin City University
Denkowski M, Lavie A (2011) Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation, 30-31 July 2011, Edinburgh, UK, pp 85–91
Doherty S, O'Brien S, Carl M (2010) Eye tracking as an MT evaluation technique. Mach Transl 24:1–13. doi:10.1007/s10590-010-9070-9
Duchowski AT (2007) Eye tracking methodology: theory and practice. Springer, London
Dunteman GH (1989) Principal components analysis. Sage, Newbury Park
Epskamp S, Cramer AOJ, Waldrop LJ, Schmittmann VD, Borsboom D (2012) qgraph: network visualizations of relationships in psychometric data. J Stat Softw. doi:10.18637/jss.v048.i04
Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI conference on human factors in computing systems, 27 Apr–2 May 2013, Paris, France, pp 439–448. doi:10.1145/2470654.2470718
Guerberof A (2014) The role of professional experience in post-editing from a quality and productivity perspective. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 51–76
Immonen S, Mäkisalo J (2010) Pauses reflecting the processing of syntactic units in monolingual text production and translation. Hermes 44:45–62
Jolliffe IT (2002) Principal components analysis, 2nd edn. Springer, New York
Just MA, Carpenter PA (1980) A theory of reading: from eye fixation to comprehension. Psychol Rev 87:329–354
Kahneman D (1973) Attention and effort. Prentice-Hall, New Jersey
Kirschner PA (2002) Cognitive load theory: implications of cognitive load theory on the design of learning. Learn Instr 12:1–10. doi:10.1016/S0959-4752(01)00014-7
Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the seventh workshop on statistical machine translation, 7-8 June 2012, Montréal Canada, pp 181–190
Koponen M, Aziz W, Ramos L, Specia L (2012) Post-editing time as a measure of cognitive effort. In: Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012), 28 Oct 2012. San Diego, USA, pp 11–20
Krings HP (2001) Repairing texts: empirical investigations of machine translation post-editing processes. In: Koby GS (ed). Kent State University Press, Kent, Ohio
Kumpulainen M (2015) On the operationalisation of 'pauses' in translation process research. Translation & Interpreting 7(1):47–58. doi:ti.106201.2015.a04
Lacruz I, Shreve G (2014) Pauses and cognitive effort in post-editing. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 246–272
Lacruz I, Shreve G, Angelone E (2012) Average pause ratio as an indicator of cognitive effort in post-editing: a case study. In: Proceedings of the workshop on post-editing technology and practice, AMTA 2012, San Diego, 28 Oct 2012, pp 29–38
Lacruz I, Denkowski M, Lavie A (2014) Cognitive demand and cognitive effort in post-editing. In: Proceedings of the third workshop on post-editing technology and practice, AMTA 2014, 22–26 Oct 2014. Vancouver, BC Canada, pp 73–84
Meara P, Buxton B (1987) An alternative to multiple choice vocabulary tests. Lang Test 4:142–154
Mitchell L (2015) Community post-editing of machine-translated user-generated content. PhD Thesis, Dublin City University
Moorkens J, O'Brien S (2015) Post-editing evaluations: Trade-offs between novice and professional participants. In: Durgar El-Kahlout I, Özkan M, Sánchez-Martínez F, Ramírez-Sánchez G, Hollowood F,

Way A (eds) Proceedings of European Association for Machine Translation (EAMT) 2015, Antalya, pp 75–81
Moorkens J, O'Brien S, da Silva IAL, de Lima Fonseca NB, Alves F (2015) Correlations of perceived post-editing effort with measurements of actual effort. Mach Transl 29:267–284. doi:10.1007/s10590-015-9175-2
Moray N (1967) Where is capacity limited? A survey and a model. Acta Psychol 27:84–92. doi:10.1016/0001-6918(67)90048-0
O'Brien S (2006) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Lang Cult 7:1–21. doi:10.1556/Acr.7.2006.1.1
O'Brien S (2011) Towards predicting post-editing productivity. Mach Transl 25:197–215. doi:10.1007/s10590-011-9096-7
O'Donnell RD, Eggemeier FT (1986) Workload assessment methodology. In: Boff KR, Kaufman L, Thomas JP (eds) Handbook of perception and human performance, cognitive processes and performance, vol 2. Wiley, New York, pp 42-41–42-49
Olsen A (2012) The Tobii I-VT fixation filter—algorithm description. http://acuity-ets.com/downloads/Tobii%20I-VT%20Fixation%20Filter.pdf. Accessed 03 Dec 2016
Paas F (1992) Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. J Educ Psychol 84:429–434
Paas F, Tuovinen JE, Tabbers H, van Gerven PWM (2003) Cognitive load measurement as a means to advance cognitive load theory. Educ Psychol 38(1):63–71. doi:10.1207/S15326985EP3801_8
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localization context. Prague Bull Math Linguist 93:7–16
Read J (2007) Second language vocabulary assessment: current practices and new directions. Int J Engl Stud 7:105–126
Schilperoord J (1996) It's about time: temporal aspects of cognitive processes in text production. Rodopi, Amsterdam
Snijders TAB, Bosker R (1999) Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage, London
Temnikova I (2010) A cognitive evaluation approach for a controlled language post-editing experiment. In: Proceedings of the seventh international conference on language resources and evaluation, 19–21 May 2010. Valetta, Malta, pp 3485–3490
Unsworth N, Heitz RP, Schrock JC, Engle RW (2005) An automated version of the operation span task. Behav Res Methods 37:498–505
Unsworth N, Redick TS, Heitz RP, Broadway JM, Engle RW (2009) Complex working memory span tasks and higher-order cognition: a latent-variable analysis of the relationship between processing and storage. Memory 17:635–654
van Gog T, Kester L, Nievelstein F, Giesbers B, Paas F (2009) Uncovering cognitive processes: different techniques that can contribute to cognitive load research and instruction. Comput Hum Behav 25:325–331. doi:10.1016/j.chb.2008.12.021
van der Maaten LJP, Hinton GE (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579–2605
Vieira LN (2014) Indices of cognitve effort in machine translation post-editing. Mach Transl 28:187–216. doi:10.1007/s10590-014-9156-x
Vieira LN (2016) Cognitive effort in post-editing of machine translation: evidence from eye movements, subjective ratings, and think-aloud protocols. PhD Thesis, Newcastle University